SIFTER: Search Services for Digital Libraries

CLIOH: Cultural digital Library Indexing Our Heritage

Mathew J. Palakal

SIFTER Research Team

CLIOH Research Team

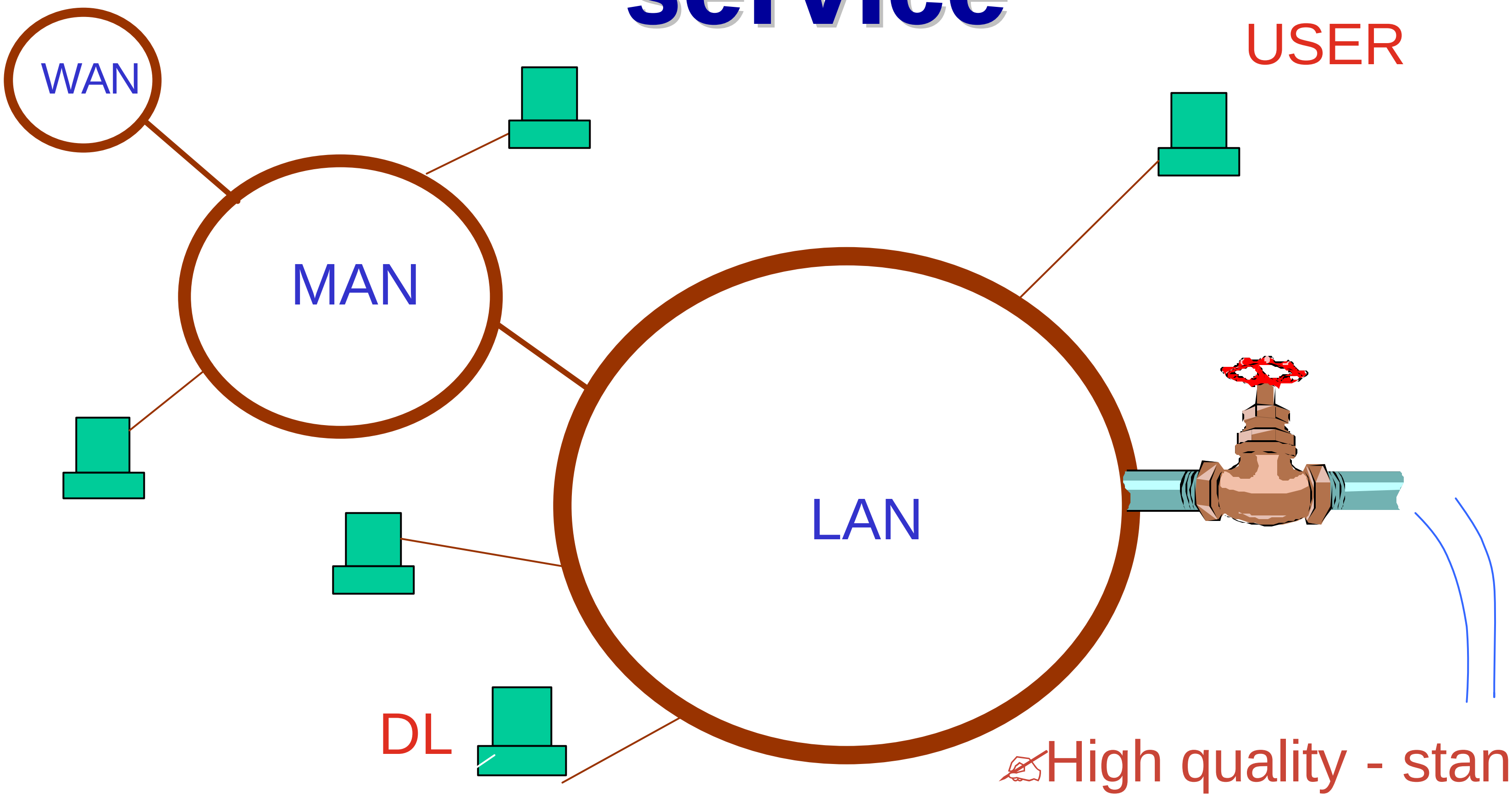
Indiana University Purdue University Indianapolis
Indiana University Bloomington

DLI/IMLS/NSDL PI Meeting, Portland, July 18, 2002

Search

- Information searching is a basic necessity …
  - Critical to the usefulness of a digital library
- Information available through a digital library may actually come from many different sources (both historical and recent)
- Users may need access to multiple digital libraries – distributed across the globe

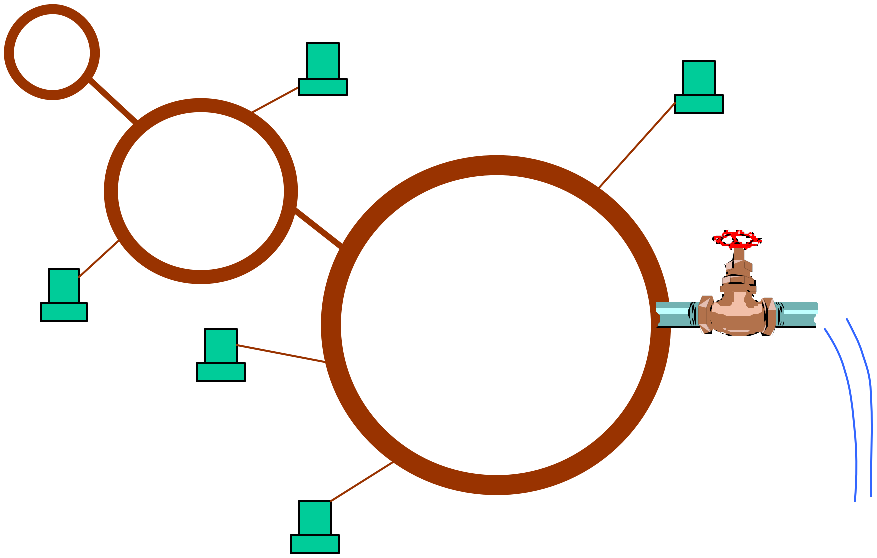
Search as a “utility” service



- ✍High quality - standards
- ✍Persistent - always on
- ✍Robust – scalable

- ✍Smart – “demand aware”

Effective Search Service



- Organization
  - Aggregation
  - Representation
  - Classification
- Matching
  - Delivery media & devices: (customization)
  - Users’ interests: (query & profile personalization)
- Presentation & interaction
  - Prune, cluster, rank, format, visualize

Key Challenges

1. Data Diversity
   - Diverse sources
   - Numerous formats
   - Heterogeneous content
2. Dynamic Environment
   - Content drift
   - Quality change
3. User needs
   - User’s demands are context-sensitive
   - User’s interest vary and may change over time

Rising to the Challenge

?

SIFTER

Smart Information Filtering Technologies for Electronic Resources

Developing algorithms and systems that utilize both IR and AI approaches

Problem 1: Data Diversity

- Need to identify document semantics: labels, terms and concepts
- Need to identify associations among concepts, terms or labels

Data Diversity: SIFTER Solutions

- **Representation**:
  - Use of thesauri
  - Algorithms to convert data elements to efficiently computable structures
- **Classification**:
  - Use of comprehensive classification schemes
  - Algorithms to cluster or classify to higher level representations

Problem 2: Dynamic Sources

- Local users -> local vocabularies and functions
- New “vocabularies” are introduced and they need to be discovered

Dynamic Sources: SIFTER Solutions

- Distributed knowledge and functions using multi-agent architecture
- Vocabulary discovery based on discriminatory power
- Classification scheme generation and ongoing replenishment

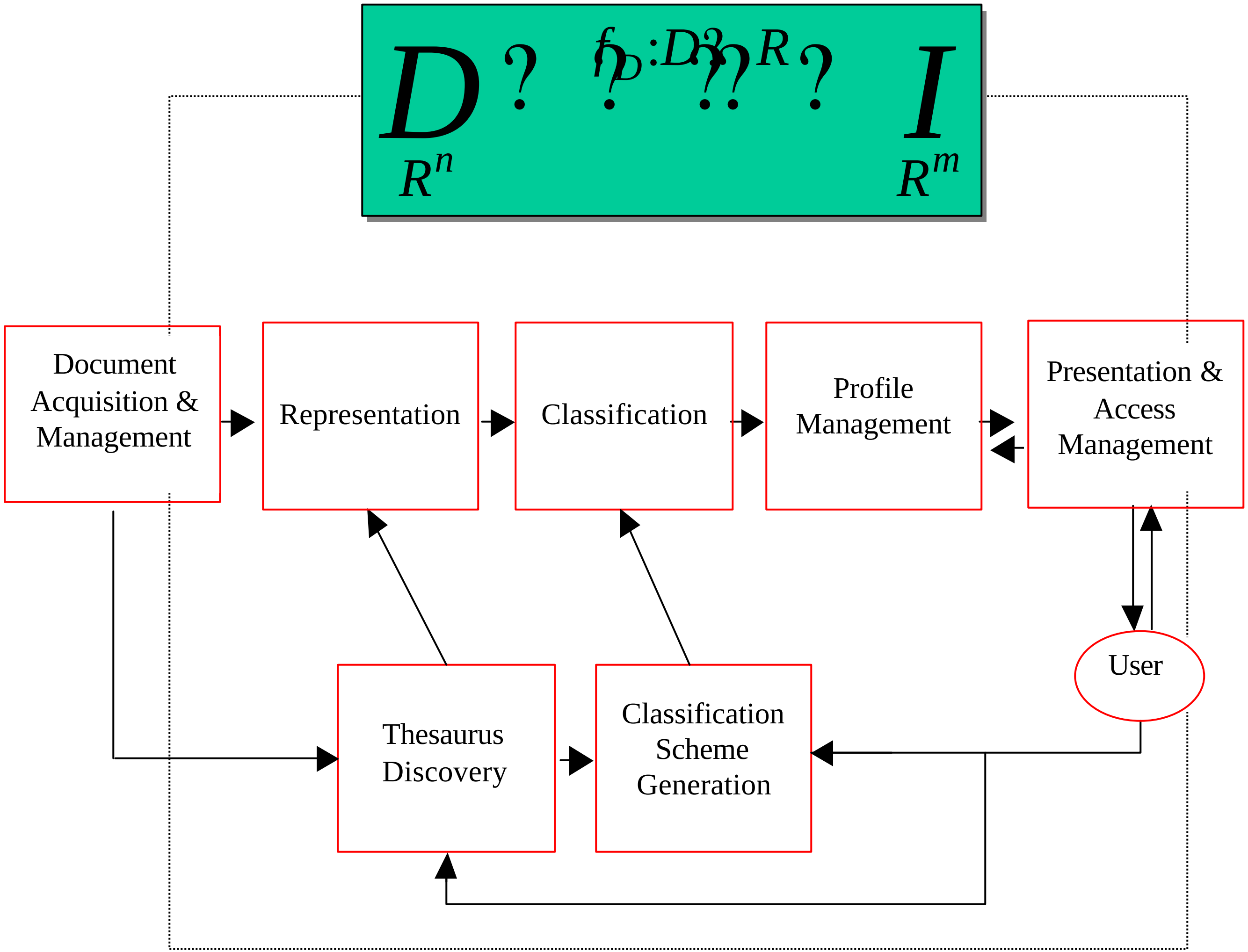
Problem 3: User Need

- Interest information usually covers a small subset of the universe of topics (a profile), therefore service has to be “personalized”
- Identification and capturing of profile cannot rely directly and exclusively on the user
- Interest may not be constant over topics and can change gradually or rapidly

User Need: SIFTER Solution

- User profile modeling: capture interest in a representation that supports topic “exploration” and reduces user involvement
- Promote control and convenient modification
- Detect “interest shifts”
- Support model and domain visualizations

Modeling the Information Filtering Process



D ? $f_D : D$?? R ? I

R^n R^m

Document Acquisition & Management

Representation

Classification

Profile Management

Presentation & Access Management

User

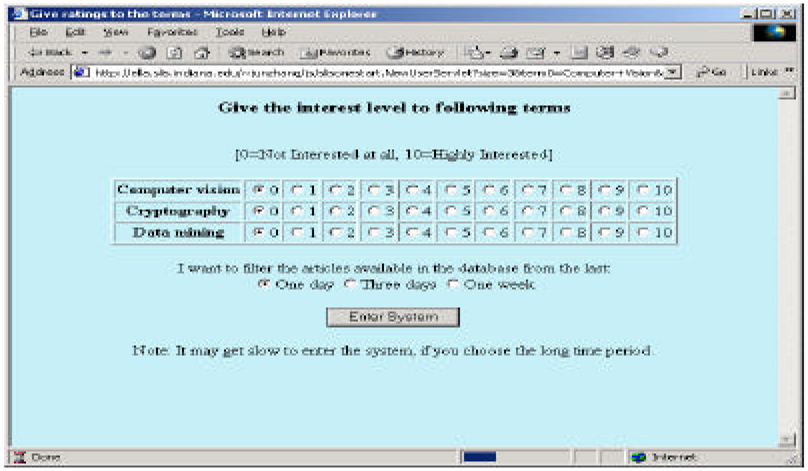
Thesaurus Discovery

Classification Scheme Generation

J. Mostafa, S. Mukhopadhyay, M. Palakal and W. Lam, A Multi-level Approach to Intelligent Information Filtering: Model, System, and Evaluation, *ACM Transactions on Information Systems*, Vol. 15, No. 4, pp. 368-399, 1997

Capturing User’s Interest

- Explicit (topics), rating content, and user behavior


Give ratings to the terms - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://ella.slis.indiana.edu/~junzhang/js/slisonestart.NewUserServlet?size=3&termID=Computer+Vision&
Go Links
Done
Internet

Give the interest level to following terms

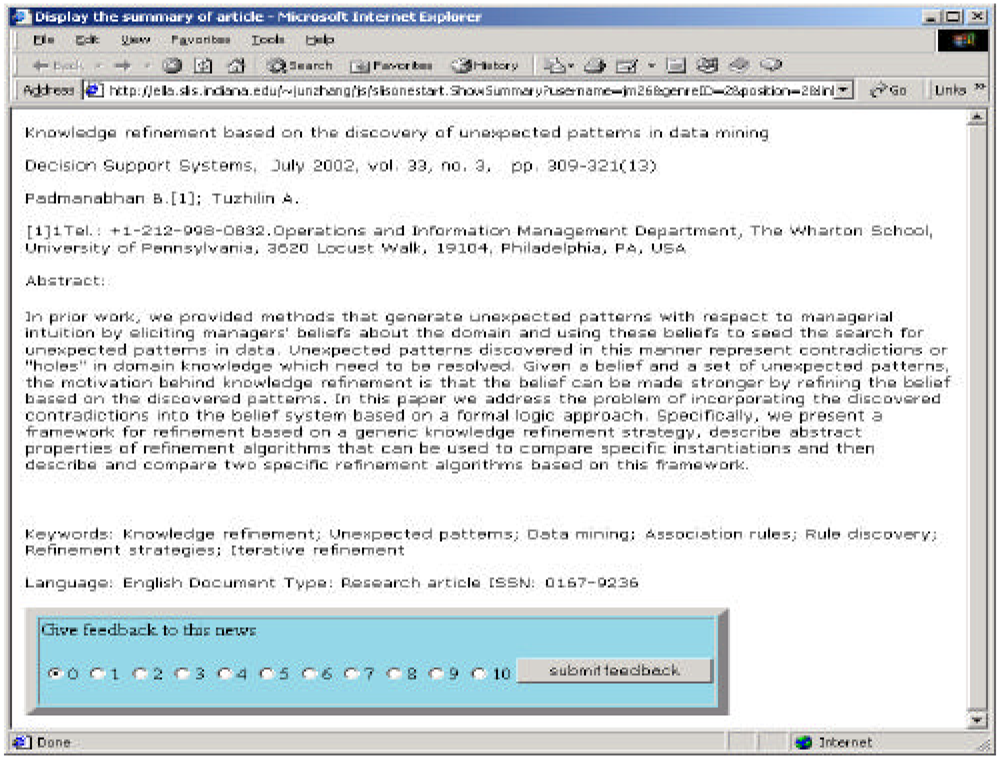
[0=Not Interested at all, 10=Highly Interested]

Computer vision	● 0	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7	○ 8	○ 9	○ 10
Cryptography	● 0	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7	○ 8	○ 9	○ 10
Data mining	● 0	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7	○ 8	○ 9	○ 10

I want to filter the articles available in the database from the last:
● One day ○ Three days ○ One week

Enter System

Note: It may get slow to enter the system, if you choose the long time period.


Display the summary of article - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://ella.slis.indiana.edu/~junzhang/js/slisonestart.ShowSummary?username=jm26&genreID=2&position=2&in
Go Links
Done
Internet

Knowledge refinement based on the discovery of unexpected patterns in data mining

Decision Support Systems, July 2002, vol. 33, no. 3, pp. 309-321(13)

Padmanabhan B.[1]; Tuzhilin A.

[1]1Tel.: +1-212-998-0832.Operations and Information Management Department, The Wharton School, University of Pennsylvania, 3620 Locust Walk, 19104, Philadelphia, PA, USA

Abstract:

In prior work, we provided methods that generate unexpected patterns with respect to managerial intuition by eliciting managers' beliefs about the domain and using these beliefs to seed the search for unexpected patterns in data. Unexpected patterns discovered in this manner represent contradictions or "holes" in domain knowledge which need to be resolved. Given a belief and a set of unexpected patterns, the motivation behind knowledge refinement is that the belief can be made stronger by refining the belief based on the discovered patterns. In this paper we address the problem of incorporating the discovered contradictions into the belief system based on a formal logic approach. Specifically, we present a framework for refinement based on a generic knowledge refinement strategy, describe abstract properties of refinement algorithms that can be used to compare specific instantiations and then describe and compare two specific refinement algorithms based on this framework.

Keywords: Knowledge refinement; Unexpected patterns; Data mining; Association rules; Rule discovery; Refinement strategies; Iterative refinement

Language: English Document Type: Research article ISSN: 0167-9236

Give feedback to this news

● 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10 submit feedback

Lam, W. & Mostafa, J. "Modeling User Interest Shift Using a Bayesian Approach".
***Journal of the American Society for Information Science & Technology* , 52(5), 416-429, 2001**

Representation & Classification

- Representation:
  - Use of thesauri
  - Algorithms to convert data stream to efficiently computable structures
- Classification:
  - Algorithms to cluster or classify to higher level representations

M. Palakal, Y. Mulong, S. Mukhopadhyay, The Effects of Dynamic Clustering in Information Filtering, *Information Processing and Management*, 2002 (in preparation).

Automated Approaches

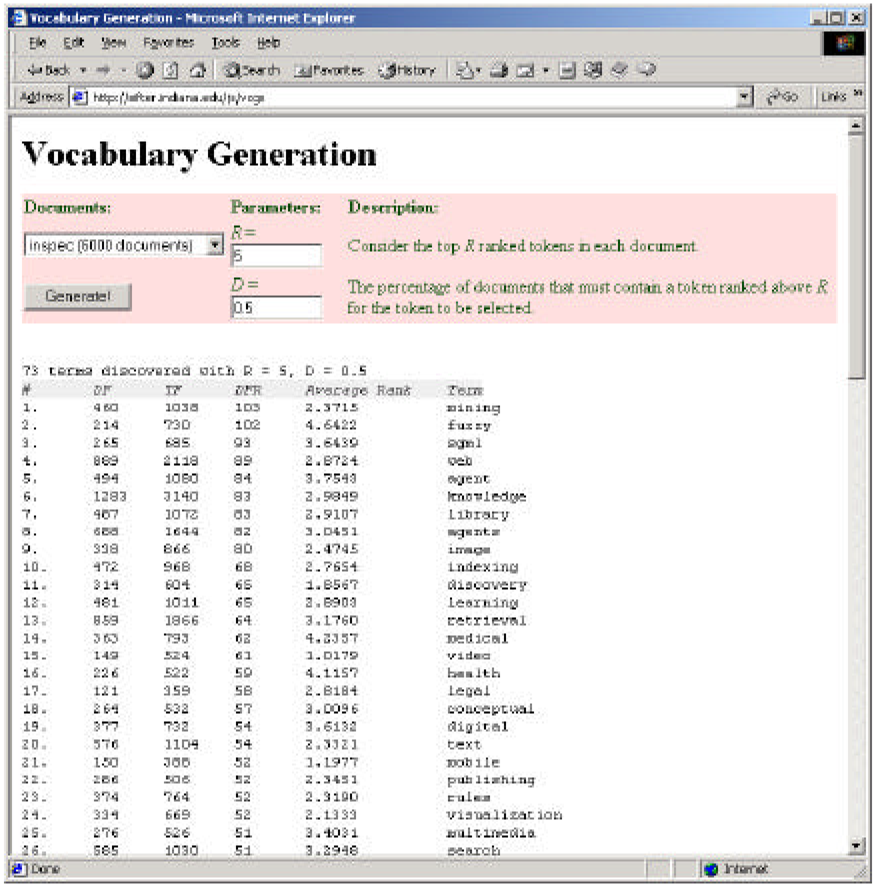
- Learned from existing classification results – used PUBMED for training
- Developed algorithms for vocabulary and association discovery

MeSH Classes
Cell Adhesion
Cell Communication
Cell Death
Cell Movement
Cell Survival
Endocytosis
Antibody Formation
Autoimmunity
Immunocompromised Host
Cytotoxicity Immunologic
Immune Tolerance
Immunity Cellular
Regeneration
Evolulution
Complement Activation

Automatically Produced Classes
Cell, Binding
Cell, Adhesion, Growth, Antigen
Communication, Death
Apoptois
Migration
Production, Motility
Tolerance
Virus
Endocytosis, Receptor
Antibody, Serum
Autoimmune
Tumor
Immunocompromised, Infected
Cytotoxic
Immune, Cell, Response, Gene, Class
Regeneration
Evolution, DNA
Complement, Activation, Plasma, Membrane
Transplant
Muscle
Expression

Mostafa, J., & Lam, W. "Automatic Classification Using Supervised Learning in a Medical Document Filtering Application." ***Information Processing & Management*****, 36(3), 415-444, 2000**

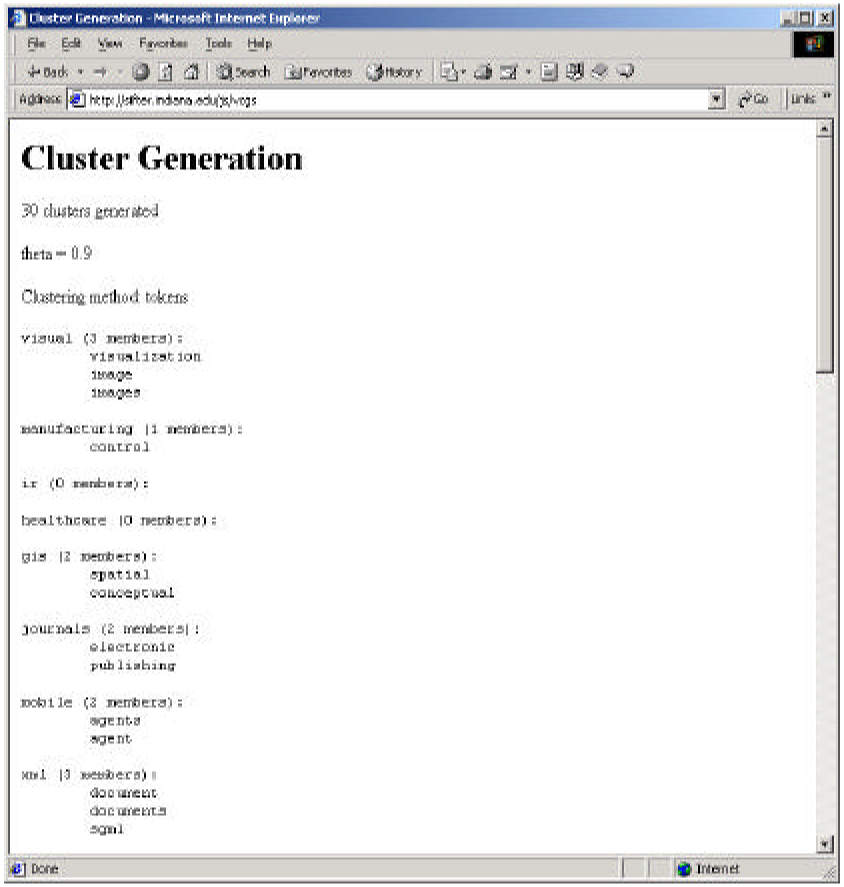
Interactive Term and Cluster Discovery


Vocabulary Generation - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://sifter.indiana.edu/js/vcgi
Go Links
Done
Internet

Vocabulary Generation

Documents:	Parameters:	Description:
inspec (6000 documents)	*R* = 5	Consider the top *R* ranked tokens in each document.
Generate!	*D* = 0.5	The percentage of documents that must contain a token ranked above *R* for the token to be selected.

```
73 terms discovered with R = 5, D = 0.5
#        DF      TF      DFR     Average Rank      Term
1.       460     1038    103     2.3715            mining
2.       214     730     102     4.6422            fuzzy
3.       265     685     93      3.6439            sgml
4.       889     2118    89      2.8724            web
5.       494     1080    84      3.7543            agent
6.       1283    3140    83      2.9849            knowledge
7.       467     1072    83      2.9107            library
8.       688     1644    82      3.0451            agents
9.       338     866     80      2.4745            image
10.      472     968     68      2.7654            indexing
11.      314     604     65      1.8567            discovery
12.      481     1011    65      2.8903            learning
13.      859     1866    64      3.1760            retrieval
14.      363     793     62      4.2357            medical
15.      149     324     61      1.0179            video
16.      226     522     59      4.1157            health
17.      121     359     58      2.8184            legal
18.      264     532     57      3.0096            conceptual
19.      377     732     54      3.6132            digital
20.      576     1104    54      2.3321            text
21.      150     388     52      1.1977            mobile
22.      286     506     52      2.3451            publishing
23.      374     764     52      2.3190            rules
24.      334     669     52      2.1333            visualization
25.      276     526     51      3.4031            multimedia
26.      585     1030    51      3.2948            search
```


Cluster Generation - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://sifter.indiana.edu/js/vcgs
Go Links
Done
Internet

Cluster Generation

30 clusters generated

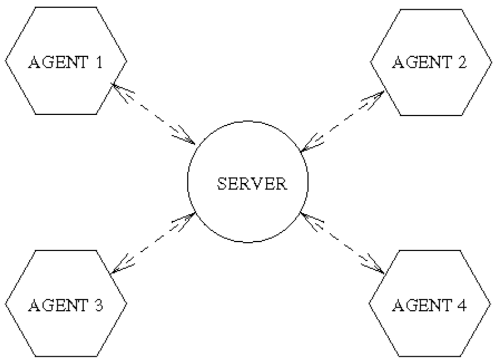
theta = 0.9

Clustering method: tokens

```
visual (3 members):
        visualization
        image
        images

manufacturing (1 members):
        control

ir (0 members):

healthcare (0 members):

gis (2 members):
        spatial
        conceptual

journals (2 members):
        electronic
        publishing

mobile (2 members):
        agents
        agent

xml (3 members):
        document
        documents
        sgml
```

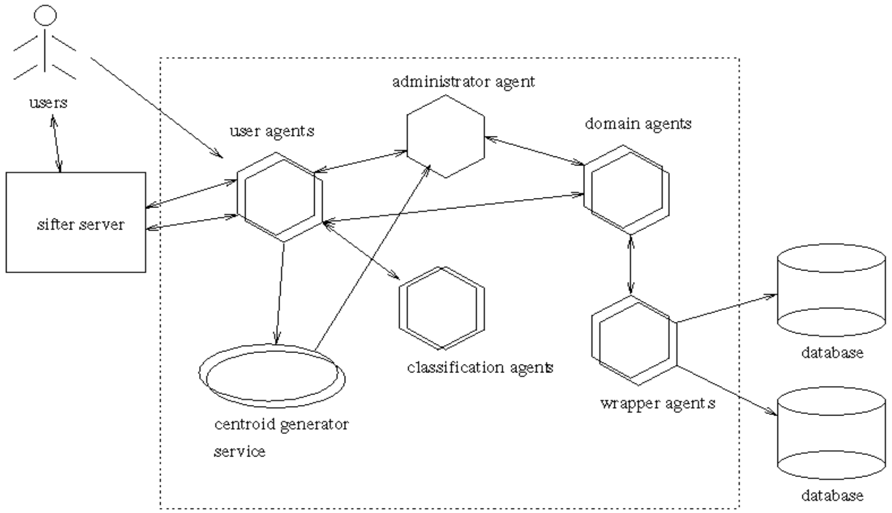
Diverse Sources: Distributed Services

D-SIFTER



Distributed knowledge & Local functionality

SIFTER-II



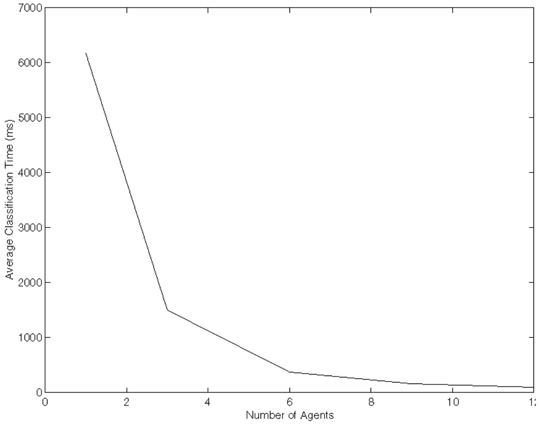
Distributed knowledge & Distributed functionality

Raje, R., Qiao, M., Mukhopadhyay, S., Palakal, M., & J. Mostafa, J. “Homogeneous Agent-based Distributed Information Filtering'', *Cluster Computing,* 2002 (in press)

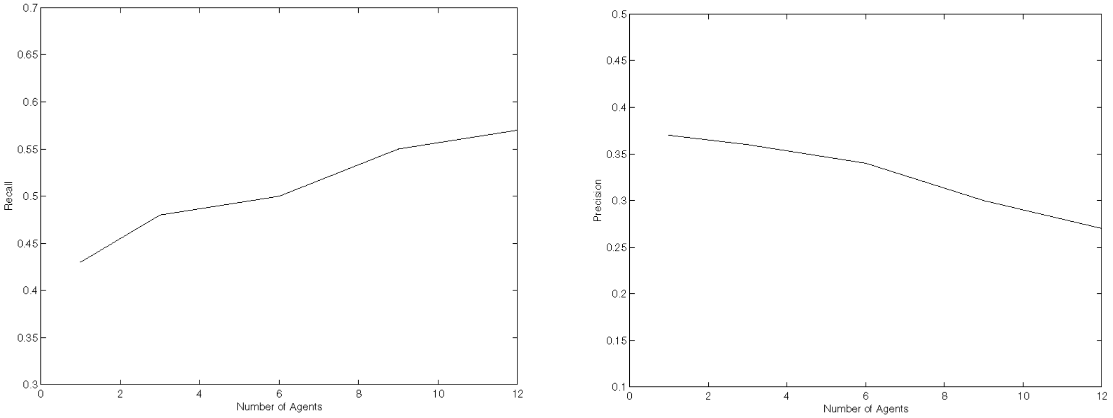
Raje, R., Qiao, M., Mukhopadhyay, S., Palakal, M., & J. Mostafa, SIFTER-II: A Heterogeneous Agent Society for Information Filtering, *Proceedings of ACM Symposium on Applied Computing, SAC'01*, pp. 121-123, Las Vegas, Nevada, 2002.

Evaluation of Distributed Filtering

- Progressively larger number of users -> larger agent community -> time to classify decreases
- Growing user community -> increasing number of agents -> precision suffers moderately but recall improves
- With increasing number of “user agents” classification efficiency improves



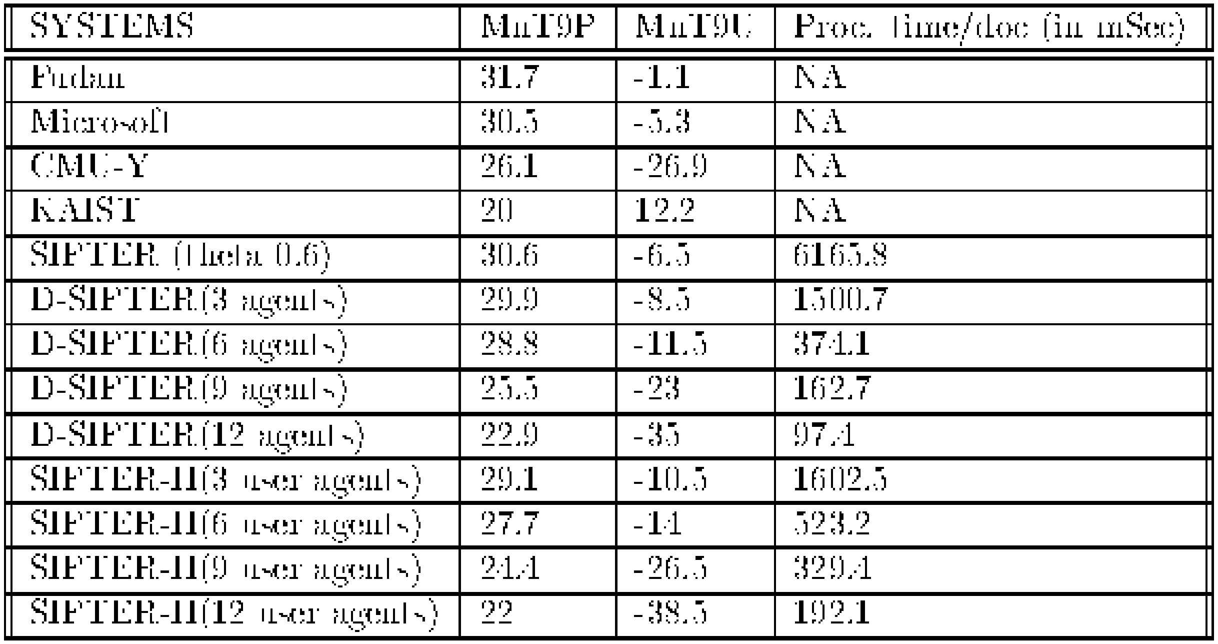
Processing Time



Filtering Performance

SIFTER vs. the best in TREC

- TREC 2000 Filtering Track OHSUMED collection was used
- 293,856 documents in the test set
- 4967 topics (include OHSU and MeSH topics)
- Evaluated BOTH effectiveness and efficiency



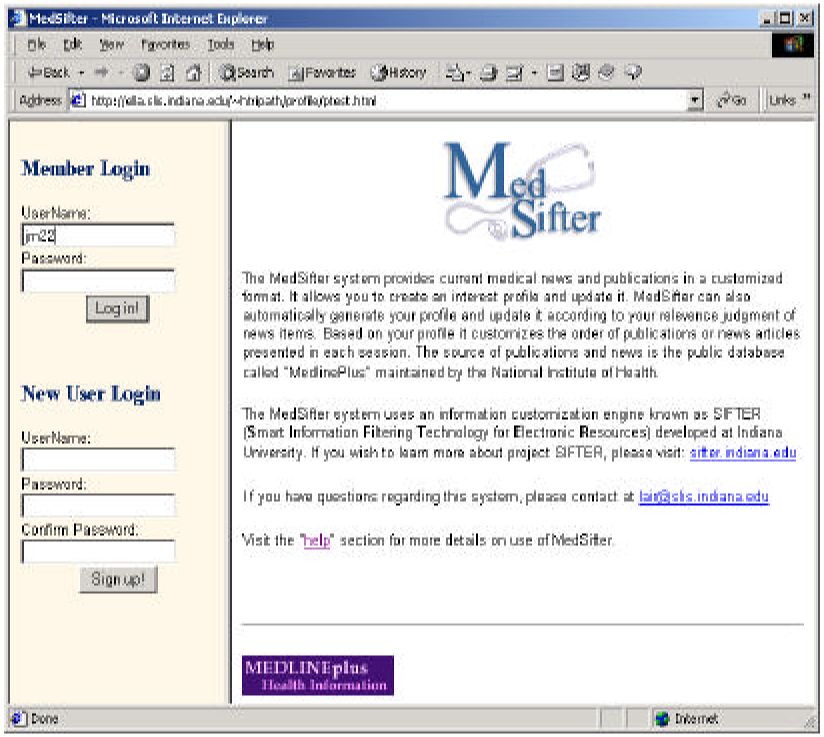
SYSTEMS	MnT9P	MnT9U	Proc. time/doc (in mSec)
Fudan	31.7	-1.1	NA
Microsoft	30.5	-5.3	NA
CMU-Y	26.1	-26.9	NA
KAIST	20	12.2	NA
SIFTER (theta 0.6)	30.6	-6.5	6165.8
D-SIFTER(3 agents)	29.9	-8.5	1500.7
D-SIFTER(6 agents)	28.8	-11.5	374.1
D-SIFTER(9 agents)	25.5	-23	162.7
D-SIFTER(12 agents)	22.9	-35	97.4
SIFTER-II(3 user agents)	29.1	-10.5	1602.5
SIFTER-II(6 user agents)	27.7	-14	523.2
SIFTER-II(9 user agents)	24.4	-26.5	329.4
SIFTER-II(12 user agents)	22	-38.5	192.1

S. Mukhopadhyay, S. Peng, M. Qiao, R. Raje, J. Mostafa, and M. Palakal, Distributed Multi-Agent Information Filtering, *ACM Transactions on Information Systems*, 2002 (pending review).

Diverse Formats

- Developing systems for health news (text), scholarly research publications (text), music (audio) and cultural information (all major formats)
  - MedSIFTER
  - TuneSIFTER
  - BioSIFTER
  - ViewFinder (CLIOH)

Systems for Different Data Formats: MedSIFTER


MedSifter - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://ella.slis.indiana.edu/~htripath/profile/ptest.html
Go Links
Med Sifter
Done
Internet

Member Login

UserName: jm22

Password:

Log in!

New User Login

UserName:

Password:

Confirm Password:

Sign up!

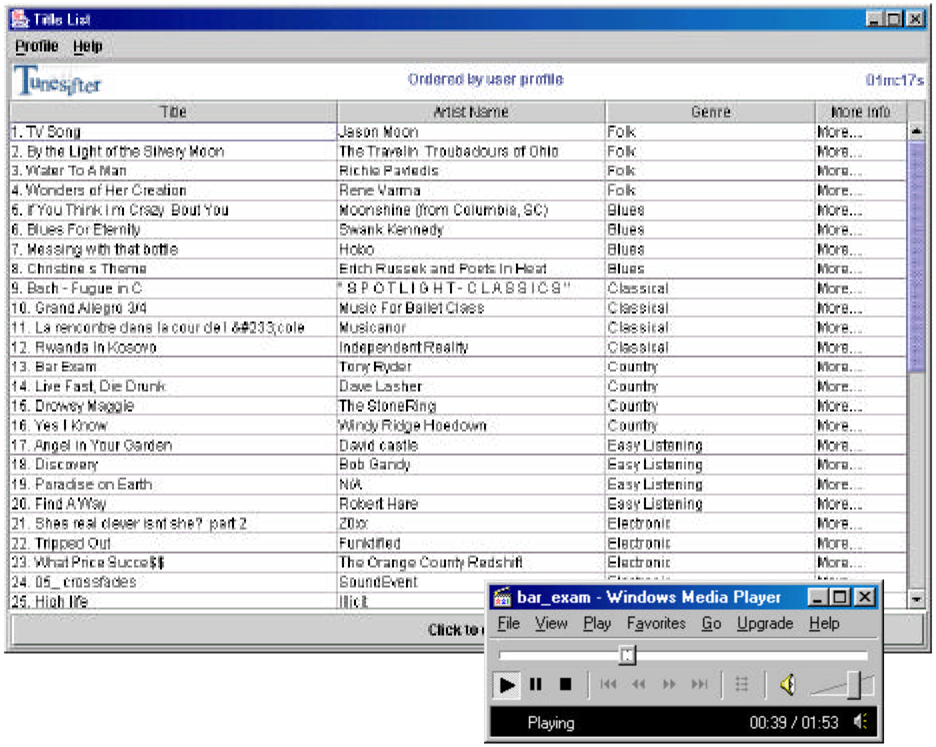
The MedSifter system provides current medical news and publications in a customized format. It allows you to create an interest profile and update it. MedSifter can also automatically generate your profile and update it according to your relevance judgment of news items. Based on your profile it customizes the order of publications or news articles presented in each session. The source of publications and news is the public database called "MedlinePlus" maintained by the National Institute of Health.

The MedSifter system uses an information customization engine known as SIFTER (Smart Information Filtering Technology for Electronic Resources) developed at Indiana University. If you wish to learn more about project SIFTER, please visit: sifter.indiana.edu

If you have questions regarding this system, please contact at lair@slis.indiana.edu

Visit the "help" section for more details on use of MedSifter.

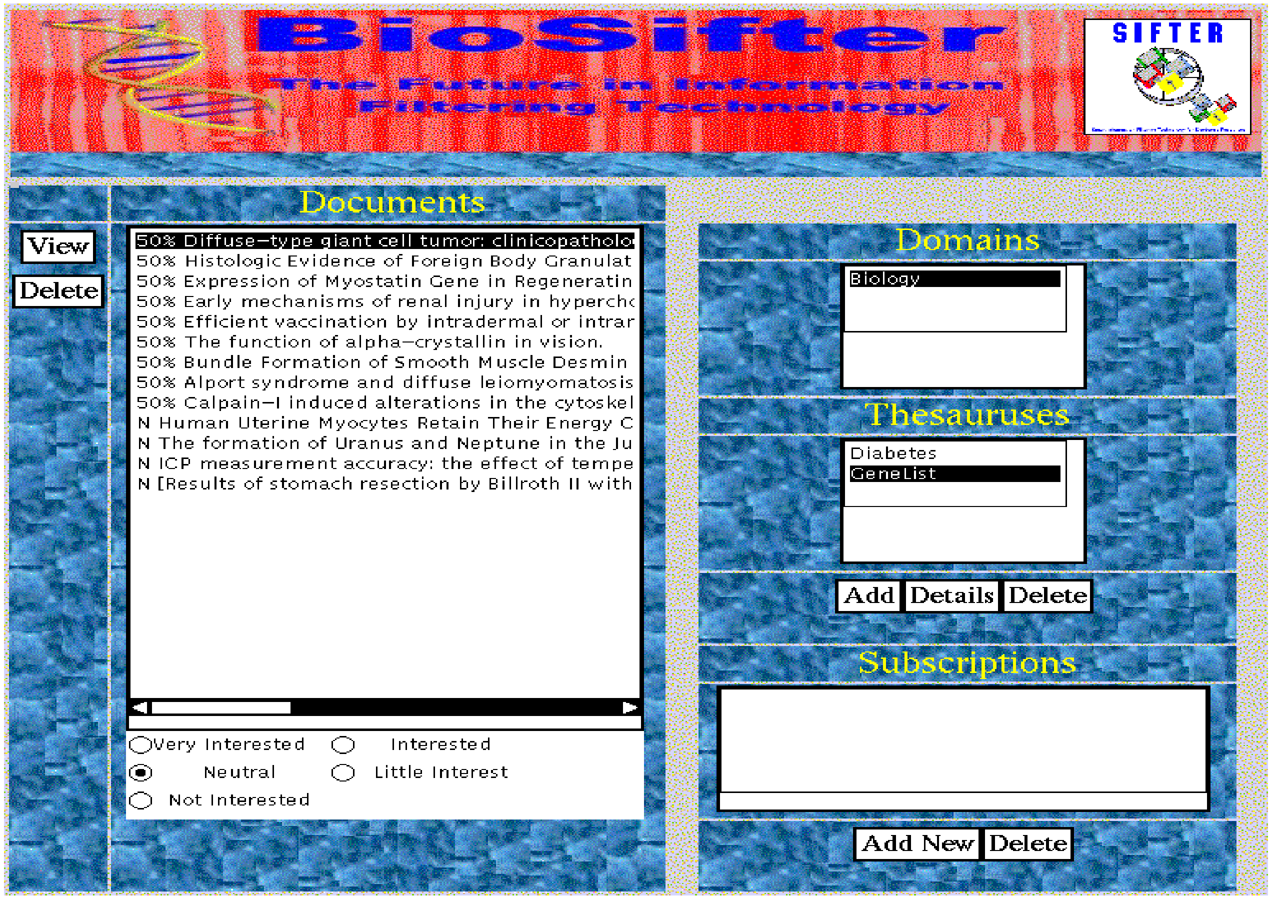
MEDLINEplus Health Information



Title	Artist Name	Genre	More Info
1. TV Song	Jason Moon	Folk	More...
2. By the Light of the Silvery Moon	The Travelin Troubadours of Ohio	Folk	More...
3. Water To A Man	Richie Pavledis	Folk	More...
4. Wonders of Her Creation	Rene Varma	Folk	More...
5. If You Think I m Crazy Bout You	Moonshine (from Columbia, SC)	Blues	More...
6. Blues For Eternity	Swank Kennedy	Blues	More...
7. Messing with that bottle	Hobo	Blues	More...
8. Christine s Theme	Erich Russek and Poets In Heat	Blues	More...
9. Bach - Fugue in C	"SPOTLIGHT-CLASSICS"	Classical	More...
10. Grand Allegro 3/4	Music For Ballet Class	Classical	More...
11. La rencontre dans la cour de l &#233;cole	Musicanor	Classical	More...
12. Rwanda In Kosovo	Independent Reality	Classical	More...
13. Bar Exam	Tony Ryder	Country	More...
14. Live Fast, Die Drunk	Dave Lasher	Country	More...
15. Drowsy Maggie	The StoneRing	Country	More...
16. Yes I Know	Windy Ridge Hoedown	Country	More...
17. Angel in Your Garden	David castle	Easy Listening	More...
18. Discovery	Bob Gandy	Easy Listening	More...
19. Paradise on Earth	N/A	Easy Listening	More...
20. Find A Way	Robert Hare	Easy Listening	More...
21. Shes real clever isnt she? part 2	Z0x	Electronic	More...
22. Tripped Out	Funkified	Electronic	More...
23. What Price Succe$$	The Orange County Redshift	Electronic	More...
24. 05_ crossfades	SoundEvent	[illegible]	[illegible]
25. High life	Illicit		

http://ella.slis.indiana.edu/~htripath/profile/ptest.html

Systems for Different Data Formats: BioSIFTER



50% Diffuse–type giant cell tumor: clinicopatholo
50% Histologic Evidence of Foreign Body Granulat
50% Expression of Myostatin Gene in Regeneratin
50% Early mechanisms of renal injury in hyperchc
50% Efficient vaccination by intradermal or intrar
50% The function of alpha–crystallin in vision.
50% Bundle Formation of Smooth Muscle Desmin
50% Alport syndrome and diffuse leiomyomatosis
50% Calpain–I induced alterations in the cytoskel
N Human Uterine Myocytes Retain Their Energy C
N The formation of Uranus and Neptune in the Ju
N ICP measurement accuracy: the effect of tempe
N [Results of stomach resection by Billroth II with

○ Very Interested ○ Interested
● Neutral ○ Little Interest
○ Not Interested

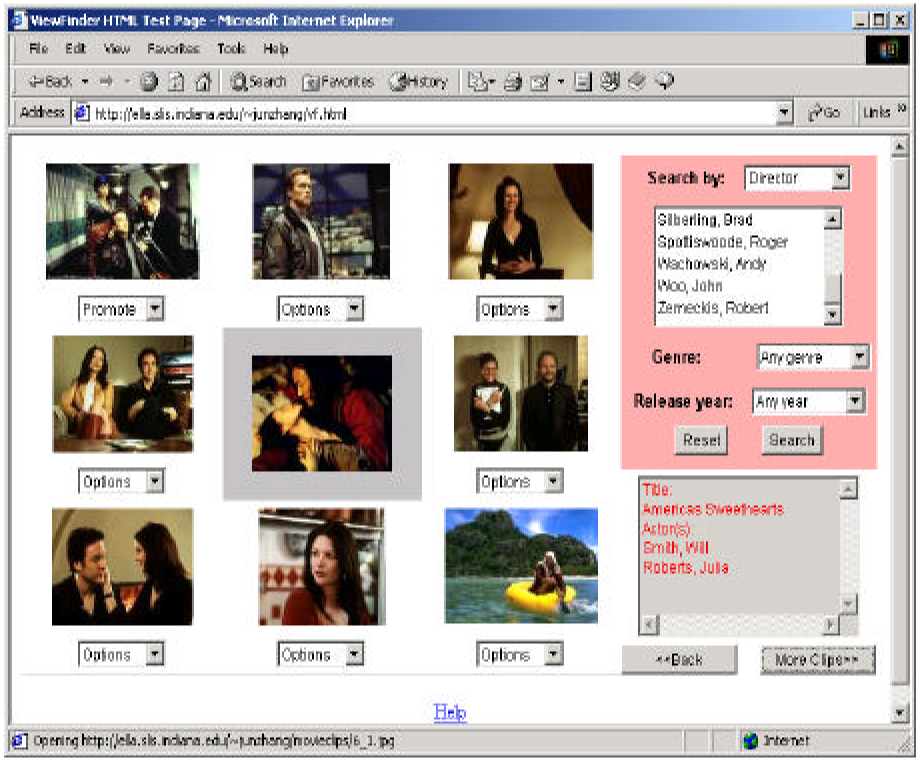
Domains: Biology

Thesauruses: Diabetes, GeneList

- Text Data
- Sequence Data
- Structural Data

M. Palakal, S. Mukhopadhyay, J. Mostafa, R. Raje, M. N'Cho, and S.K. Mishra, An Intelligent Biological Information Management System, *Bioinformatics*, 2002, (in press).

Systems for Different Data Formats: CLIOH



ViewFinder HTML Test Page - Microsoft Internet Explorer

Address http://ella.slis.indiana.edu/~junzhang/vf.html

Search by: Director

Silberling, Brad
Spottiswoode, Roger
Wachowski, Andy
Woo, John
Zemeckis, Robert

Genre: Any genre

Release year: Any year

Title:
Americas Sweethearts
Actor(s):
Smith, Will
Roberts, Julia

Opening http://ella.slis.indiana.edu/~junzhang/movieclips/6_1.jpg

ViewFinder

- Video Data
- Audio Data

Beyond Current Challenges

- Cross-format, cross-language, and cross-domain information synthesis in real-time.
- Distributed DLs with both Data & Services
- Integrating Web Services with Multi-agent Searching

Acknowledgment

This project was funded in part by:

- The National Science Foundation (DLI-II and ITR I)
- Eli Lilly & Co., Indianapolis, Indiana

SIFTER Team:

- Mathew Palakal, Snehasis Mukhopadhyay, Javed Mostafa, Rajeev Raje
- Students: Mulong Yu, Matthew Stephens, Mingyaung Qiao, Shengquan Peng, Luz Quiroga, John Fieber, Vijay Vij